Datasets

This list of datasets below has been compiled for the convenience of the students for reference purposes only. The list is provided as it is, without warranty of any kind. It is students' responsibility to wisely select and carefully check the datasets according to the guidelines and discussion outlined in the lectures prior to use. Note that the list below is not exhaustive, and it is the responsibility of the students to conduct thorough research to identify (additional) suitable datasets for the Group Mini Project.

Please note that some datasets may only be obtainable upon request from the dataset authors. In these cases, it is recommended to contact the authors as soon as possible to avoid any delays this process might create.

HateMM Dataset
HateMM: A Multi-Modal Dataset for Hate Video Classification (2023)
Paper: https://arxiv.org/abs/2305.03915v1
Repository: https://zenodo.org/records/7799469
Creative Commons Attribution 4.0 International

OLID: Offensive Language Identification Dataset
Predicting the Type and Target of Offensive Posts in Social Media (2020)
Paper: https://arxiv.org/pdf/1902.09666v2
We are checking the license information for this dataset, and will update the information soon.

Hate Speech and Offensive Language Dataset
Automated Hate Speech Detection and the Problem of Offensive Language (2017)
Paper: https://ojs.aaai.org/index.php/ICWSM/article/view/14955/14805
Repository: https://github.com/t-davidson/hate-speech-and-offensive-language
MIT License

ThreatGram 101: Extreme Telegram Replies Data with Threat Levels Dataset
Exploring Multi-Level Threats in Telegram Data with AI-Human Annotation: A Preliminary Study (2023)
Paper: https://ieeexplore.ieee.org/document/10459792
Repository: https://data.mendeley.com/datasets/tm9s68vgxd/1
CC BY 4.0

Implicit Hate Dataset
Latent Hatred: A Benchmark for Understanding Implicit Hate Speech (2021)
Paper: https://aclanthology.org/2021.emnlp-main.29.pdf
Repository: https://github.com/SALT-NLP/implicit-hate
MIT License

Civil Comments: Jigsaw Unintended Bias in Toxicity Classification Dataset
Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification (2019)
Paper: https://dl.acm.org/doi/10.1145/3308560.3317593
Repository: https://huggingface.co/datasets/google/civil_comments
Creative Commons Zero v1.0 Universal

Hatemoji Dataset
Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate (2022)

Paper: https://arxiv.org/abs/2108.05921
Repository: https://github.com/HannahKirk/Hatemoji

More datasets are available (but not limited to) at: https://github.com/leondz/hatespeechdata/tree/master

HatEval Dataset
SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (2019)
Paper: https://aclanthology.org/S19-2007.pdf
Repository: https://github.com/msang/hateval
Creative Commons CC-BY-NC-4.0

MetaHate Dataset
MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection (2024)
Paper: https://ojs.aaai.org/index.php/ICWSM/article/view/31445/33605
Repository: https://github.com/palomapiot/metahate?tab=readme-ov-file
Creative Commons Attribution Non Commercial Share Alike 4.0

HateSpeech Dataset
Hate Speech Dataset from a White Supremacy Forum (2018)
Paper: https://aclanthology.org/W18-51.pdf
Repository: https://github.com/Vicomtech/hate-speech-dataset?tab=License-1-ov-file
Creative Commons Attribution-ShareAlike 3.0 Spain License

ETHOS: multi-labEl haTe speecH detectiOn dataSet
ETHOS: an Online Hate Speech Detection Dataset (2021)
Paper: https://link.springer.com/epdf/10.1007/s40747-021-00608-2
Repository: https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset
GNU GPLv3

MLMA Hate Speech Dataset
Multilingual and Multi-Aspect Hate Speech Analysis (2019)
Paper: https://aclanthology.org/D19-1474.pdf
Repository: https://github.com/HKUST-KnowComp/MLMA_hate_speech
MIT license

CONAN: COunter NArratives through Nichesourcing Dataset
CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech (2019)
Paper: https://aclanthology.org/P19-1271.pdf
Repository: https://github.com/marcoguerini/CONAN
“These resources can be used for research purposes and cannot be redistributed. Please cite the corresponding publication if you use any dataset.”

Knowledge-grounded hate countering Dataset
Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech (2021)
Paper: https://aclanthology.org/2021.findings-acl.79.pdf
Repository: https://github.com/marcoguerini/CONAN

“These resources can be used for research purposes and cannot be redistributed. Please cite the corresponding publication if you use any dataset.”

Multitarget CONAN Dataset
Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech (2021)
Paper: https://aclanthology.org/2021.acl-long.250.pdf
Repository: https://github.com/marcoguerini/CONAN
“These resources can be used for research purposes and cannot be redistributed. Please cite the corresponding publication if you use any dataset.”

DIALOCONAN Dataset
Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering (2022)
Paper: https://aclanthology.org/2022.emnlp-main.549.pdf
Repository: https://github.com/marcoguerini/CONAN
“These resources can be used for research purposes and cannot be redistributed. Please cite the corresponding publication if you use any dataset.”

Hate Speech and Offensive Language Dataset
Automated Hate Speech Detection and the Problem of Offensive Language (2017)
Paper: https://ojs.aaai.org/index.php/ICWSM/article/view/14955/14805
Repository: https://github.com/t-davidson/hate-speech-and-offensive-language
Note: the repository is no longer actively maintained.
MIT License

TOXIGEN Dataset
ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection (2022)
Paper: https://arxiv.org/abs/2203.09509
Repository: https://github.com/microsoft/ToxiGen

Multimodal Sarcasm Detection Dataset: MUStARD Dataset
Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper) (2019)
Note: The dataset is compiled from popular TV shows including Friends, The Golden Girls, The Big Bang Theory, and Sarcasmaholics Anonymous.
Paper: https://aclanthology.org/P19-1455/
Repository: https://github.com/soujanyaporia/MUStARD

MUStARD++ Dataset
A Multimodal Corpus for Emotion Recognition in Sarcasm (2022)
Note: The dataset is compiled from popular TV shows including Friends, The Golden Girls, The Big Bang Theory, and Sarcasmaholics Anonymous.
Paper: https://arxiv.org/abs/2206.02119
Repository: https://github.com/cfiltnlp/MUStARD_Plus_Plus

Online Misogyny Dataset

An expert annotated dataset for the detection of online misogyny (2021)
Paper: https://aclanthology.org/2021.eacl-main.114/
Repository: https://github.com/ellamguest/online-misogyny-eacl2021

CAD: Contextual Abuse Dataset
Introducing CAD: the Contextual Abuse Dataset (2021)
Paper: https://aclanthology.org/2021.naacl-main.182/
Repository: https://github.com/dongpng/cad_naacl2021, https://zenodo.org/records/4881008

A Multilingual Dataset of Racial Stereotypes Dataset
A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads (2023)
Paper: https://aclanthology.org/2023.findings-eacl.51/
Note: The annotated dataset will be available for research purposes upon request, together with the complete set of annotation guidelines.

News-Headlines-Dataset-For-Sarcasm-Detection
Sarcasm detection using news headlines Dataset (2023)
Paper: https://www.sciencedirect.com/science/article/pii/S2666651023000013#fn5
Repository: https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection
https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection/data

iSarcasm Dataset
iSarcasm: A Dataset of Intended Sarcasm (2019)
Paper: https://arxiv.org/abs/1911.03123v2
Repository: https://github.com/dmbavkar/iSarcasm
Note from the dataset authors: While we only make the tweet IDs public, we maintain and are happy to provide the following for research purposes, under an agreement that protects the privacy of our users: tweet texts; for each sarcastic tweet, an explanation given by its authors as to why the tweet is sarcastic; for each sarcastic tweet, a rephrase given by its author that conveys the same message non-sarcastically.

Large-Scale Hate Speech Detection Dataset
Large-Scale Hate Speech Detection with Cross-Domain Transfer (2022)
Paper: https://aclanthology.org/2022.lrec-1.238/
Repository: https://github.com/avaapm/hatespeech

HateXplain Dataset
HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection (2021)
Paper: https://ojs.aaai.org/index.php/AAAI/article/view/17745
Repository: https://github.com/hate-alert/HateXplain

ConvAbuse Dataset
ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI (2021)
Paper: https://aclanthology.org/2021.emnlp-main.587/
Repository: https://github.com/amandacurry/convabuse

SWAD: Swear Words Abusiveness Dataset

Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media (2020)
Paper: https://aclanthology.org/2020.lrec-1.765/
Repository: https://github.com/dadangewp/SWAD-Repository

World of Warcraft Cyberbullying Dataset
Detecting cyberbullying in online communities (2016)
Paper: https://core.ac.uk/download/pdf/301369744.pdf
Repository: http://ub-web.de/research/

League of Legends Cyberbullying Dataset
Detecting cyberbullying in online communities (2016)
Paper: https://core.ac.uk/download/pdf/301369744.pdf
Repository: http://ub-web.de/research/

MultiOFF: Multimodal Meme Dataset
Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text (2020)
Paper: https://aclanthology.org/2020.trac-1.6.pdf
Repository: https://drive.google.com/drive/folders/1hKLOtpVmF45IoBmJPwojgq6XraLtHmV6

ALONE: AdoLescents ON twittEr Dataset
ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter (2020)
Paper: https://arxiv.org/pdf/2008.06465.pdf
Repository: https://zenodo.org/records/3608352
Note from the dataset authors: we make this dataset available upon request to the authors, and researchers will be required to sign an agreement to use it only for research purposes and without public dissemination.

MMHS150K Dataset
Exploring Hate Speech Detection in Multimodal Publications (2019)
Paper: https://arxiv.org/pdf/1910.03814.pdf
Repository: https://gombru.github.io/2019/10/09/MMHS/

HarMeme Dataset
Detecting Harmful Memes and Their Targets (2021)
Paper: https://arxiv.org/abs/2110.00413
Repository:https://github.com/di-dimitrov/harmeme

OASST1: OpenAssistant Conversations Dataset
OpenAssistant Conversations – Democratizing Large Language Model Alignment (2023)
Paper:https://proceedings.neurips.cc/paper_files/paper/2023/file/949f0f8f32267d297c2d4e3ee10a2e7e-Paper-Datasets_and_Benchmarks.pdf
Repository: https://github.com/LAION-AI/Open-Assistant
Apache-2.0 license

FairPrism Dataset
FairPrism: Evaluating Fairness-Related Harms in Text Generation (2023)
Repository: https://github.com/microsoft/FairPrism
Paper: https://aclanthology.org/2023.acl-long.343.pdf

MIT license
Note on dataset access: reach out to the fairprism@microsoft.com to access the dataset.

HarmfulQA Dataset
Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment (2023)
Paper: https://arxiv.org/abs/2308.09662
Repository: https://github.com/declare-lab/red-instruct
Apache-2.0 license

MultiHateClip Dataset
MultiHateClip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili (2024)
Paper: https://dl.acm.org/doi/pdf/10.1145/3664647.3681521
Repository: https://github.com/social-ai-studio/multihateclip?tab=readme-ov-file

MAMI: Multimedia Automatic Misogyny Identification Dataset
SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (2022)
Paper: https://aclanthology.org/2022.semeval-1.74/
Repository:https://github.com/MIND-Lab/SemEval2022-Task-5-Multimedia-Automatic-Misogyny-Identification-MAMI-
"The data may be distributed upon request and for academic purposes only. To request the datasets, please fill out the following form: https://forms.gle/AGWMiGicBHiQx4q98
After submitting the required info, participants will have a link to a folder containing the datasets in a zip format (trial, training and development) and the password to uncompress the files."

HateComments Dataset
Hateful Comment Detection and Hate Target-Type Prediction for Video Comments (2023)
Paper: https://dl.acm.org/doi/pdf/10.1145/3583780.3615260
Repository: https://drive.google.com/file/d/1EUbWDUokv1CYkWKlwByUC6yIuBGUw2MN/

The "Call me sexist, but" sexism dataset
Paper: https://ojs.aaai.org/index.php/ICWSM/article/view/18085/17888
Repository: https://search.gesis.org/research_data/SDN-10.7802-2251?doi=10.7802
Website flickers

Hateful Meme Dataset
The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes (2020)
Paper: https://arxiv.org/abs/2005.04790
Repository: https://github.com/facebookresearch/mmf/tree/main/projects/hateful_memes
https://ai.meta.com/tools/hatefulmemes/
To acquire the data, you will need to register at DrivenData's Hateful Memes Competition and download data from the challenge's download page. MMF won't be able to automatically download the data since you manually need to agree to the licensing terms.
BSD License